Nucleic Acids Research, 2004, Vol. 32, Database issue ***D165–D167***
DOI: 10.1093/nar/gkh062

Update of NUREBASE: nuclear hormone receptor functional genomics

David Ruau, Jorge Duarte, Tarik Ourjdal, Guy Perrière[1], Vincent Laudet and Marc Robinson-Rechavi*

Laboratoire de Biologie Moléculaire de la Cellule, CNRS UMR 5161, Ecole Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07, France and [1]Laboratoire de Biométrie et Biologie Evolutive, CNRS UMR 5558, Université Claude Bernard, Lyon 1, 43 boulevard du 11 Novembre 1918, 69622 Villeurbanne Cedex, France

Received September 12, 2003; Accepted September 25, 2003

ABSTRACT

Nuclear hormone receptors are an abundant class of ligand-activated transcriptional regulators, found in varying numbers in all animals. Based on our experience of managing the official nomenclature of nuclear receptors, we have developed NUREBASE, a database containing protein and DNA sequences, reviewed protein alignments and phylogenies, taxonomy and annotations for all nuclear receptors. New developments in NUREBASE include explicit declaration of alternative transcripts of each gene, and expression data for human and mouse nuclear receptors. The core of NUREBASE is reviewed, and it is completed by NUREBASE_DAILY, automatically updated every 24 h. All information on accessing and installing NUREBASE may be found at http://www.ens-lyon.fr/LBMC/laudet/nurebase/nurebase.html.

INTRODUCTION

Nuclear hormone receptors are one of the most abundant classes of transcriptional regulators in metazoans (1). They function as ligand-activated transcription factors, and include receptors for hydrophobic molecules such as steroid hormones, retinoic acids and thyroid hormones (2). As nuclear receptors bind small molecules which can easily be modified by drug design, and control functions associated with major pathologies (cancer, osteoporosis, diabetes, etc.), they are important pharmacological targets. A large number of nuclear receptors, referred to as 'nuclear orphan receptors', have also been identified by homology but have no identified natural ligand, and may allow the development of new drugs (3,4).

The importance of nuclear receptors has prompted the accumulation of data from a great diversity of fields of research. The aim of NUREBASE is to present an integrated database, centralizing up-to-date information about nuclear receptors for the specialist and the non-specialist (5). NUREBASE also serves as the implementation of the official nomenclature of nuclear receptors (6), of which one of us (V.L.) is in charge. Release 04 (September 2003) of NUREBASE contains 523 nuclear receptor protein entries without redundancy, from 115 metazoan species. The sequences are grouped into 30 nested 'families' [see (5)], corresponding to levels of nomenclature, each with an alignment and a phylogeny. NUREBASE also contains 771 nuclear receptor cDNA entries (see next section). These reviewed data compose the core of NUREBASE. It is completed by a daily automatic update procedure, which identifies and classifies nuclear receptors, and adds them to NUREBASE_DAILY.

The main way to access NUREBASE is through the PBIL website (7), which allows users to perform queries centered either on sequences or on families. Queries on sequences are based on criteria common to most retrieval systems: entry names, accession numbers, keywords, taxonomic data, bibliographic references, etc. In the case of queries centered on families, it is possible to retrieve all gene families that are shared by a given set of taxa and that are not present in a second set of taxa (i.e. vertebrates but not human). Utilities are also available on the PBIL server to visualize and handle the alignments and trees integrated in NUREBASE. Alignments can be simply displayed with a coloring scheme in an HTML document or can be viewed with the JalView applet (http://www.ebi.ac.uk/~michele/jalview/contents.html). Likewise, trees can be displayed as clickable pictures or through the ATV (A Tree Viewer) applet (8).

All information on using NUREBASE may be found at http://www.ens-lyon.fr/LBMC/laudet/nurebase/nurebase.html.

DNA SEQUENCES AND ALTERNATIVE TRANSCRIPTS

In the first version of NUREBASE (5), EMBL (9) DNA entries were linked 'as is' to the annotated NUREBASE protein entries. DNA entries are now classified and annotated. An important point is that DNA sequences are limited to the exact coding sequence: they do not contain introns or non-coding flanking regions. This choice was made because: (i) most experimentalists studying nuclear receptors use RT–PCR on mRNA, and the coding sequence is what is relevant to

*To whom correspondence should be addressed at present address: Joint Center for Structural Genomics, UCSD, 9500 Gilman Drive, La Jolla, CA 92093-0527, USA. Tel: +1 858 646 3100; Fax: +1 858 646 3171; Email: marc.robinson@ens-lyon.fr

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors

design PCR primers, or compare a new cDNA sequence with the previously available sequences; (ii) more and more protein sequences are associated with DNA from genome sequencing projects, and it is usually not relevant to use a complete chromosome sequence as a nuclear receptor entry. Of course the EMBL accession number remains available, so the user can follow a link to the original DNA sequence, and access information about non-coding regions. The NUREBASE DNA entry corresponding to a protein has the same name, but with zero as a separator instead of an underscore: the DNA entry for protein THA_HOMS1 is THA0HOMS1. These DNA entries have the same keywords as the protein entry, in this case 'Thyroid hormone; Triiodothyronine; T3', allowing direct queries on DNA entries.

One gene can code for several proteins, via mechanisms such as alternative splicing, alternative promoters, etc. (10). These alternative proteins can play major functional roles in nuclear receptors (1). Because the family system underlying NUREBASE is based on protein phylogeny, in which each homologous gene should be represented by only one leaf (or OTU), we decided to keep one 'standard' protein sequence per nuclear receptor gene, defined as having the canonical A/B-C-D-E(-F) domain organization. This protein is associated with the exact cDNA that encodes it, as explained in the previous paragraph, plus an unlimited number of alternative cDNAs encoding alternative proteins. The detailed procedure to identify alternative cDNAs will be described elsewhere (J. Duarte *et al.*, in preparation), but briefly we define them as having regions of $\geqslant$99% local identity, yet <90% global identity (counting gaps as differences). In NUREBASE, alternative cDNAs are identified by names with the same format as standard cDNAs, but with a digit superior to 0 as a separator: in addition to THA0HOMS1, for the THA gene, there are alternative cDNAs THA1HOMS1 to THA7HOMS1. As the function of the encoded proteins may be different (indeed some alternative forms of THA do not bind thyroid hormone) (11), we do not associate functional keywords (such as ligand) with these sequences. Specific keywords such as 'alternative' allow queries on this important feature of nuclear receptor genes in NUREBASE. It should be noted that we only identify in NUREBASE alternative transcripts that (i) modify the coding sequence, and (ii) have been cloned and sequenced as cDNAs, i.e. we do not search for alternative transcripts from expressed sequence tag (EST) data. This ensures maximum relevance and minimum false positives in the alternative cDNA information presented.

EXPRESSION DATA

When molecular biologists characterize a gene, one of the first and most important questions is 'when and where is it expressed?' Yet few bioinformatic resources allow just this question to be answered in a straightforward manner, because of difficulties in obtaining this information (data mining), and in representing it. We believe nuclear receptors are a good model to tackle these problems, because of the vast amount of information available, and the well-known structure of the superfamily. We have identified for human and mouse (*Mus musculus*) all nuclear receptor ESTs from dbEST (12). These were classified according to 56 tissues or organs (such as 'brain' or 'liver'), three development stages ('fetal', 'infant'/'newborn' and 'adult'), and 'tumor' if relevant. Only when there were at least two EST entries with the same information (same organ, same stage, same tumoral state) was the information retained, to minimize false positives. This information was then organized in keywords of the form 'organ tumor stage', where each field may be absent, so examples of keywords are: 'brain tumor adult', 'brain fetal', 'tumor adult', 'brain', etc. Thanks to the use of the wildcard, this allows all possible queries: all genes expressed in tumors ('*tumor*'), or those expressed in brain tumors only ('brain tumor*'), or all those expressed in brain ('brain*'), etc. We are presently working on an automation of this procedure, as well as its extension to other species and other sources of expression data.

Preliminary analysis shows that results are biologically and clinically relevant: RXRs (NR2B), which are ubiquitous heterodimer partners for other nuclear receptors, are found in the largest variety of organs, while the PNR (photo-receptor specific nuclear receptor; NR2E3) is found only in the eye. Concerning tumors, a key role of RXRβ (NR2B2) is suggested since it is the only nuclear receptor found in all gonadic and related tumors (cervix, uterus, testis, prostate and breast). ERs (estrogen receptors; NR3A) illustrate the limitations of the EST data and stringent criteria we used: as they are expressed in many organs but at low transcription levels, no significant expression pattern is detected.

CONCLUSION

NUREBASE provides up-to-date sequence and nomenclature information about nuclear receptors, but also serves as a resource to integrate information from functional genomics, such as alternative transcripts and expression data. The capacity to query expression patterns represents an important result for nuclear receptor bioinformatics, and potentially for other applications. Future developments include complete expression data for nuclear receptors from all species, using information not only from EST sequences but also from microarrays and SAGE, and representation of genomic information, such as introns or promoters (including alternative transcripts that do not modify the protein). Moreover, a unified interface integrating NUREBASE and other resources to query evolutionary, functional and structural information about nuclear receptors is presently in development.

ACKNOWLEDGEMENTS

NUREBASE is supported by 'Appel d'offres inter-EPST Bioinformatique'.

REFERENCES

1. Laudet,V. and Gronemeyer,H. (2002) *The Nuclear Receptors FactsBook*. Academic Press, London, UK.
2. Escriva,H., Delaunay,F. and Laudet,V. (2000) Ligand binding and nuclear receptor evolution. *Bioessays*, **22**, 717–727.
3. Gustafsson,J.A. (1999) Seeking ligands for lonely orphan receptors. *Science*, **284**, 1285–1286.
4. Kliewer,S.A., Lehmann,J.M. and Willson,T.M. (1999) Orphan nuclear receptors: shifting endocrinology into reverse. *Science*, **284**, 757–760.
5. Duarte,J., Perrière,G., Laudet,V. and Robinson-Rechavi,M. (2002) NUREBASE: database of nuclear hormone receptors. *Nucleic Acids Res.*, **30**, 364–368.

6. Nuclear Receptors Nomenclature Committee (1999) A unified nomenclature system for the nuclear receptor superfamily. *Cell*, **97**, 161–163.
7. Perrière,G., Combet,C., Penel,S., Blanchet,C., Thioulouse,J., Geourjon,C., Grassot,J., Charavay,C., Gouy,M., Duret,L. *et al.* (2003) Integrated databanks access and sequence/structure analysis services at the PBIL. *Nucleic Acids Res.*, **31**, 3393–3399.
8. Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.
9. Stoesser,G., Baker,W., van den Broek,A., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V., Lopez,R. *et al.* (2003) The EMBL nucleotide sequence database: major new developments. *Nucleic Acids Res.*, **31**, 17–22.
10. Graveley,B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, **17**, 100–107.
11. Gauthier,K., Plateroti,M., Harvey,C.B., Williams,G.R., Weiss,R.E., Refetoff,S., Willott,J.F., Sundin,V., Roux,J.P., Malaval,L. *et al.* (2001) Genetic analysis reveals different functions for the products of the thyroid hormone receptor α locus. *Mol. Cell. Biol.*, **21**, 4748–4760.
12. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST—database for 'expressed sequence tags'. *Nature Genet.*, **4**, 332–333.